International Journal of Scientific and Innovative Mathematical Research(IJSIMR)
Volume 8, Issue 4, 2020, PP 1-13
ISSN No. (Online) 2347-3142
DOI: https://doi.org/10.20431/2347-3142.0804001
www.arcjournals.org

Detecting Outliers using R Package in Fitting Data with Linear and Nonlinear Regression Models

Manimannan G[1*], M. Salomi[2], R. Lakshmi Priya[3], Saranraj R.[4]

[1]Assistant Professor, Department of Statistics, TMG College of Arts and Science, Chennai.
[2]Assistant Professor, Department of Statistics, Madras Christian College, Chennai
[3]Assistant Professor, Department of Statistics, Dr. Ambedkar Government Arts College, Vyasarpadi, Chennai.
[4]Assistant Professor, Department of Statistics, St. Joseph's College of Arts and Science, Cuddolore

Corresponding Author:** ***Manimannan G, *Assistant Professor, Department of Statistics, TMG College of Arts and Science, Chennai.*

Abstract:
Background
Linear and non linear regression analysis assumes scatter of data, fitting of straight line or normal distribution. An outlier is an extreme observation when the residual is larger in absolute value when compared with other observed data set. The detection of outlier can be defined as the process of detecting and subsequently excluding outliers from the given set of data. Outlier can dominate sum of square calculation and lead to misleading results. In this paper, an attempt is made to detect the outlier of linear and non linear regression models using new approach of standardized scores of detecting outliers without the use of predicted values.
Results
First describe the methods of linear and non linear regression model. The data fitted four times of regression model. Initially the original data to fit linear regression model and deduct outlier and visualize the results. Second method is to fit linear regression without outlier and visualize the data. Third method is to fit non linear regression with original data and visualize the results. Last method is to fit non linear regression for removal of outlier data and visualize the results. In both the methods only one outlier is identified and removed using standardized score. The primary data sources were collected from case sheets in a private hospital at Bangalore. In this analysis, two parameters are Age and SBP (Systolic Blood Pressure) were used. Blood pressure is measured in two types, top level Systolic blood pressure and Bottom level DBP (Diastolic Blood Pressure). Both are measured from arteries during the contraction of heart muscles of the patient.
Conclusion
The linear and nonlinear regression model fitted for original and outlier removed data. The result of linear and non linear regression for the original data is an average model. In both the models, R^2 value is less than 0.5. After removal of outlier better fit of linear and nonlinear regression model is achieved. The R^2 values are more than 0.7. The F and t statistic are significant in two models. The scatter plot clearly visualized the outlier and without outlier data for different plots. The summary statistics of both regression models results are expressed in following section. A new approach for detecting outliers without the use of predicted values have been proposed which is quite useful in detecting outliers that detects the outliers as similar to residual and standardized residual method.
Keywords: *Outlier, Linear Regression, Nonlinear Regression, Summary statistics, Residual, Predicted, Standardized Analysis and Scatter Plot Visualization.*

1. INTRODUCTION

Regression analysis is one of the most widely used statistical tools for analyzing multifactor database. It is appealing because it provides a conceptually simple method for investigating functional relationships among variables. Regression analysis is concerned with the study of dependence of one variable, on one or more other variables, called the explanatory variables, with a view to estimating and or predicting the mean or average value of the former in terms of the known or fixed in values of the latter. The problem of model selection in linear regression model has received much attention in statistical literature. For a detailed study, refer to Draper and Smith, 1981 [1]. The statistical tools available for analysis of data are "regression." Theory of regression deals with prediction of one or more variables, called "dependent (response) variables" on the basis of other variables called "independent variables." The dependent variable is also called "criterion variable." For independent

variable names such as "predictor" or "explanatory" variables are also common. When a model is used to explain dependent variable in terms of independent variable it assumes a linear relationship between them and arrive at a linear regression model or otherwise a non-linear regression model. Framstad *et.al*, 1985 [2] suggest that in simple linear regression model, the difference method can be used for detecting outliers

2. DATABASE

The primary data sources were collected from the case sheets in private hospital at Bangalore. The database had many parameters like DBP, BMI, Height, Weight, Hb, RBC, ESR, etc (Manimannan G. *et.al, 2020* [3]. In this research paper, the researcher concentrates only on Age and SBP of the respondents. In this analysis two parameters namely, Age and SBP (Systolic Blood Pressure) were used. Blood pressure is measured in two types namely, top level Systolic blood pressure (SBP) and Bottom level DBP (Diastolic Blood Pressure). Both are measured from arteries during the contraction of heart muscles of the patient.

3. MODEL DESCRIPTION

3.1. Linear Regression

The equation of a straight line functional relationship of Y on X is given by Montgomery, *et al* .[4].

$$Y = \beta_0 + \beta_1 X \tag{1}$$

which is known as simple linear regression of Y on X . β_0 and β_1 are called parameters, and should be found in equation (1) means that for a given X_i , a corresponding Y_i consists of $Y = \beta_0 + \beta_1 X_i$ and an ε_i by which an observation may fall off the true regression line. On the basis of the information available from the observations β_0 and β_1 has to be estimated.

This model is given by

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{2}$$

which is called simple linear regression model . The term ε is a random variable and is called "error term". Finding β_0 and β_1 from (X_i, Y_i) $i = 1,2,..,,n$ is called estimation of the parameters. The Ordinary Least Square (OLS) methods can be used to fit the model and estimate the parameter values. The method of least square is used to estimate the parameters β_0 and β_1 that the sum of square of the difference between the observation Y the straight line is minimum from the equation (2).The least square estimators of β_0 and β_1 must satisfy $\hat{\beta}_0$ and $\hat{\beta}_1$.

The least square estimator of the intercept

$$\hat{\beta}_0 = \underline{y} - \hat{\beta}_1 \underline{x} \tag{3}$$

The least square estimation of the slope β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \tag{4}$$

By fitting simple linear regression model is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \ldots \tag{5}$$

Equation (5) gives a point estimate of mean of Y for a particular$x$. Equation (4) is the corrected sum of square of x_i .The difference between the observed value y_i and the corresponding fitted value $\hat{Y}$ is a residual. By mathematically the residual of i[th] is

$$e_i = y_i - \hat{y}_i$$

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \qquad i = 1,2,\ldots,n \tag{6}$$

is given by Bipin *et.al*.[5]. The assumption of regression analysis is as follows :

1. The relationship between response y and the regressor's is linear, at least approximately

2. The β_j are unknown parameters to be estimated from the data.
3. It is assumed that errors are normally and independently distributed with zero expectation (mean) and constant variance σ^2: $e_i \sim NID$ (Normally and Independently Distributed) $(0, \sigma^2)$. I.
4. The errors are uncorrelated.

The assumptions 3 and 4 imply the errors and independent random variables. The approximation of the models is the t or F statistics or R^2.

3.2. Non Linear Regression

Non linear regression assumes that the scatter of data around the ideal curve follows a normal distribution. This assumption leads to familiar goal of regression to minimize the sum of square of vertical points and curve. Single outlier can dominate the sum of squares calculation which leads to misleading results. As suggested by Hample [6, 7] .It follows the following three steps.

1. Fit a curve using a new robust nonlinear regression method.
2. Analyze the residual of the robust fit, and determine whether one or more variable are outliers
3. Remove the outliers, and perform ordinary least square regression on the remaining data.

   Outlier values can seriously disturb the least square. In figure 1 an outlier falls from the line suggested by the rest of the data. If this plot is really an outlier then the estimate of the intercept might be incorrect and the residual mean square is blown up estimate of σ^2 [8].

Fitting of a polynomial regression model using the power of a single predictor (X) by the method of linear least square.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 {x_2}^2 + \ldots + \beta_k {x_k}^k \quad (7)$$

The polynomial regression co-efficient β_1 to k th for each degree . A second order k=2 is polynomial form. By using polynomial regression model in one variable is called as quadratic model.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 {x_2}^2$$

$$E(y) = f(x)$$

is a quadratic function using this model approximately tangled nonlinear relationship equation (7) we get that $y = X\beta + \varepsilon$, x is the explanatory variables $n * pX$ -matrix .

4. Result and Discussion

The medical data analysis will be executed in the following proposed algorithms using R Studio.

Step 1: The medical data imported from .xlsx file and attach () in R Studio and assign SBP as a dependent variable and Age as a independent variable, ie. (X_i, Y_i) $i = 1,2,\ldots,n$

Step 2: Get Summary statistics from the given data.

Step 3: To execute linear regression function with R syntax and summarize the results

Step 4: Predict the given data and get the SBP predicted values.

Step 5. To get the residuals values for the given data

Step 6: To get the standardized values for the given data.

Step 7. To execute scale function of linear regression for the variable Age and get the Standardized values for given data. Scale function of each element of those values is obtained by subtracting and dividing by standard deviation

Step 8. Repeat step 7 for SBP and get standardized values for the given data

Step 9: To Visualize the results for linear regression with help of different scatter plot. Manimannan G. *et.al.* [9]

The above algorithm execute the results of linear regression of Predicted value, Standardized value, Residuals, Scale function value of Age and SBP values. (In Table 1 to 3, Figure 1). The results table and figures show the extreme in the 14th row is highlighted as outlier. In Visualization part, models are fitted and highlighted in the plots. Residuals vs, Fitted, Normal QQ plot, Scale – location and Residuals vs. Leverage. All the four plots shows that the 14 person as an outlier.

The summary statistics shows that the minimum age as 17 and maximum age is 69, the minimum SBP is 110 and maximum SBP as 260. The fitted linear regression model is y = 0.9327x + 100.39, R^2 = 0.4898. In this model R^2 value is less than 0.5 and this model is average. Subsequently the outlier is removed and executes the algorithm step 1 to step 9. Cook's distance is used in linear regression analysis to find influential outliers in a set of predictor variable. It is a way to identify points that negatively affects the regression model. The plot is a combination of each given data leverage and residual values, the higher leverage and residuals, the higher the Cook's. The F and t statistic are highly significant between two variables.

Table1. *Descriptive Statistics (n = 45)*

Variables	Age	SBP
Mean	44.42	141.82
Meadian	45.00	142.82
Mode	36	120
Standard Deviation	15.390	20.509
Range	52	110
Minimum	17	110
Maximum	69	220

Table2. *Linear Regression Summary output for original data*

Regression Statistics	
R Square	0.4898
Multiple R	0.478
Intercept	100.390
Slope (Age)	0.932

R Studio Linear Regression output:

Residuals:

Min 1Q Median 3Q Max

-22.102 -5.024 0.774 2.562 75.774

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 100.3908 6.8160 14.729 < 2e-16 ***

Age 0.9327 0.1452 6.425 8.79e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.82 on 43 degrees of freedom

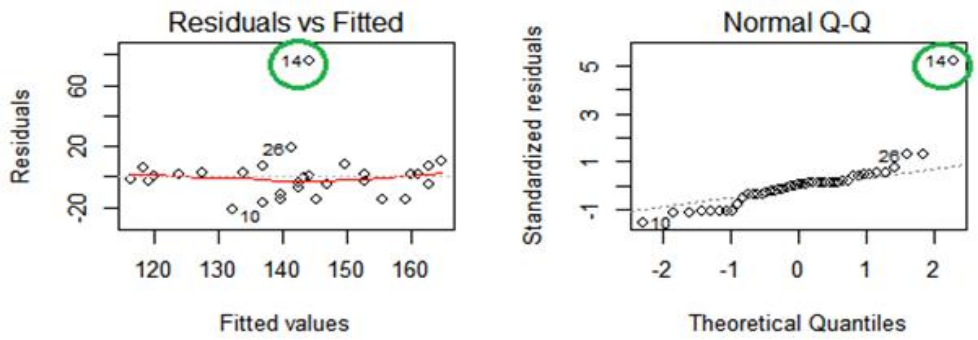
Multiple R-squared: 0.4898, Adjusted R-squared: 0.478

F-statistic: 41.29 on 1 and 43 DF, p-value: 8.791e-08

Table3. *Linear Regression of Predicted, Standardized, Residual and nonlinear Standardized and Residual Values*

S.No	Age (x_i)	SBP (y_i)	Predicted Value SBP	Standardized	Residuals	Standardized (x_i)	Standardized (y_i)
1	65	162	161.0146	0.0687	0.9854	1.3371	0.9839
2	46	142	143.2938	-0.0883	-1.2937	0.1025	0.0087
3	67	170	162.8799	0.4986	7.1201	1.4671	1.3739
4	42	124	139.5631	-1.0625	-15.5630	-0.1574	-0.8690

5	67	158	162.8799	-0.3417	-4.8799	1.4671	0.7888
6	56	154	152.6205	0.0948	1.3795	0.7523	0.5938
7	64	162	160.0819	0.1334	1.9181	1.2721	0.9839
8	56	150	152.6205	-0.1800	-2.6205	0.7523	0.3987
9	59	140	155.4185	-1.0634	-15.4180	0.9472	-0.0889
10	34	110	132.1017	-1.5165	-22.1010	-0.6772	-1.5517
11	42	128	139.5631	-0.7894	-11.5630	-0.1574	-0.6740
12	48	130	145.1591	-1.0352	-15.1590	0.2325	-0.5765
13	39	144	136.7651	0.4945	7.2349	-0.3523	0.1062
14	47	220	144.2264	5.1731	75.7735	0.1675	3.8119
15	45	138	142.3611	-0.2976	-4.3611	0.0375	-0.1864
16	47	145	144.2264	0.0528	0.7736	0.1675	0.1549
17	65	162	161.0146	0.0687	0.9854	1.3371	0.9839
18	45	135	142.3611	-0.5024	-7.3611	0.0375	-0.3327
19	17	114	116.2462	-0.1593	-2.2462	-1.7819	-1.3566
20	20	116	119.0442	-0.2141	-3.0442	-1.5869	-1.2591
21	19	124	118.1116	0.4153	5.8884	-1.6519	-0.8690
22	36	136	133.9670	0.1392	2.0330	-0.5473	-0.2839
23	50	142	147.0245	-0.3434	-5.0244	0.3624	0.0087
24	39	120	136.7651	-1.1458	-16.7650	-0.3523	-1.0641
25	21	120	119.9769	0.0016	0.0231	-1.5220	-1.0641
26	44	160	141.4284	1.2675	18.5715	-0.0274	0.8863
27	44	160	141.4284	1.2675	18.5715	-0.0274	0.8863
28	53	158	149.8225	0.5601	8.1775	0.5574	0.7888
29	63	144	159.1492	-1.0519	-15.1490	1.2072	0.1062
30	29	130	127.4383	0.1769	2.5617	-1.0021	-0.5765
31	25	125	123.7076	0.0899	1.2924	-1.2620	-0.8203
32	69	175	164.7453	0.7216	10.2547	1.5970	1.6177
33	56	154	152.6205	0.0948	1.3795	0.7523	0.5938
34	64	162	160.0819	0.1334	1.9181	1.2721	0.9839
35	36	136	133.9670	0.1392	2.0330	-0.5473	-0.2839
36	50	142	147.0245	-0.3434	-5.0244	0.3624	0.0087
37	39	120	136.7651	-1.1458	-16.7650	-0.3523	-1.0641
38	21	120	119.9769	0.0016	0.0231	-1.5220	-1.0641
39	53	158	149.8225	0.5601	8.1775	0.5574	0.7888
40	63	144	159.1492	-1.0519	-15.1490	1.2072	0.1062
41	29	130	127.4383	0.1769	2.5617	-1.0021	-0.5765
42	20	116	119.0442	-0.2141	-3.0442	-1.5869	-1.2591
43	19	124	118.1116	0.4153	5.8884	-1.6519	-0.8690
44	36	136	133.9670	0.1392	2.0330	-0.5473	-0.2839
45	50	142	147.0245	-0.3434	-5.0244	0.3624	0.0087



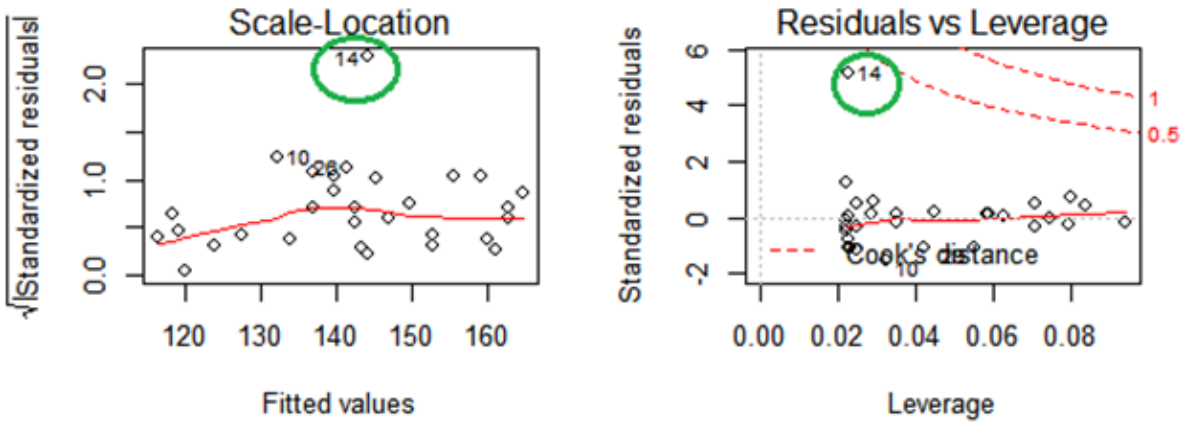


Figure1. *Scatter plot of Linear Regression models (Original Data).*

The previous algorithm again execute the results of linear regression for outlier data and get the Predicted value, Standardized value, Residuals, Scale function value of Age and SBP values. (In Table 4 to 6, Figure 2). The linear regression model improves better. In Visualization parts, models are fitted and highlighted in the plots. They are, Residuals vs, Fitted, Normal QQ plot, Scale – location and Residuals vs. Leverage. All the four plots show better results after removal of outliers.

The summary statistics shows that the minimum age is 17 and maximum age is 69, the minimum SBP is 110 and maximum SBP as 175. The fitted linear regression model is y = 0.91359x + 99.51, R^2 = 0.7091. In this model R^2 value is closer to .705 and this model is better. Subsequently the non linear regression executes the algorithm step 1 to step 9. Cook's distance is used in linear regression analysis to find influential outliers in a set of predictor variable. It is a way to identify points that negatively affects the regression model. The plot is a combination of each given data leverage and residual values, the higher leverage and residuals, the higher the Cook's. The F and t statistic are highly significant between two variables.

Table4. *Descriptive Statistics n = 44*

Variables	Age	SBP
Mean	44.363	140.0455
Median	45	141
Mode	56	162
Standard Deviation	15.562	16.88
Range	52	65
Minimum	17	110
Maximum	69	175

Table5. *Linear Regression Summary output for original data*

Regression Statistics	
R Square	0.7091
Multiple R	0.7021
Intercept	99.519
Slope (Age)	0.9134

R Studio Linear Regression output (Outlier Removed:

Residuals:

Min 1Q Median 3Q Max

-20.578 -3.194 1.922 3.996 20.287

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 99.51962 4.23951 23.47 < 2e-16 ***

Age 0.91349 0.09029 10.12 7.91e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.214 on 42 degrees of freedom

Multiple R-squared: 0.7091, Adjusted R-squared: 0.7021

F-statistic: 102.4 on 1 and 42 DF, p-value: 7.906e-13

Table6. *After Removing Outlier from Original Data (Removed 14th row)*

S. No	Age (x_i)	SBP (y_i)	Predicted Value SBP	Standardized	Residuals	Standardized (x_i)	Standardized (y_i)
1	65	162	158.8966	0.3481	3.1034	1.3260	1.3004
2	46	142	141.5403	0.0505	0.4597	0.1051	0.1158
3	67	170	160.7236	1.0451	9.2764	1.4545	1.7743
4	42	124	137.8863	-1.5250	-13.8863	-0.1519	-0.9504
5	67	158	160.7236	-0.3068	-2.7236	1.4545	1.0635
6	56	154	150.6752	0.3675	3.3248	0.7477	0.8266
7	64	162	157.9831	0.4496	4.0169	1.2618	1.3004
8	56	150	150.6752	-0.0746	-0.6752	0.7477	0.5896
9	59	140	153.4157	-1.4886	-13.4157	0.9405	-0.0027
10	34	110	130.5784	-2.2713	-20.5784	-0.6659	-1.7797
11	42	128	137.8863	-1.0857	-9.8863	-0.1519	-0.7135
12	48	130	143.3672	-1.4685	-13.3672	0.2337	-0.5950
13	39	144	135.1458	0.9734	8.8542	-0.3447	0.2342
15	45	138	140.6268	-0.2884	-2.6268	0.0409	-0.1212
16	47	145	142.4538	0.2796	2.5462	0.1694	0.2935
17	65	162	158.8966	0.3481	3.1034	1.3260	1.3004
18	45	135	140.6268	-0.6178	-5.6268	0.0409	-0.2989
19	17	114	115.0490	-0.1197	-1.0490	-1.7583	-1.5427
20	20	116	117.7895	-0.2025	-1.7895	-1.5655	-1.4243
21	19	124	116.8760	0.8081	7.1240	-1.6298	-0.9504
22	36	136	132.4053	0.3960	3.5947	-0.5374	-0.2396
23	50	142	145.1942	-0.3512	-3.1942	0.3622	0.1158
24	39	120	135.1458	-1.6652	-15.1458	-0.3447	-1.1874
25	21	120	118.7030	0.1464	1.2970	-1.5013	-1.1874
26	44	160	139.7133	2.2272	20.2867	-0.0234	1.1820
27	44	160	139.7133	2.2272	20.2867	-0.0234	1.1820
28	53	158	147.9347	1.1091	10.0653	0.5549	1.0635
29	63	144	157.0696	-1.4600	-13.0696	1.1975	0.2342
30	29	130	126.0109	0.4431	3.9891	-0.9872	-0.5950
31	25	125	122.3569	0.2957	2.6431	-1.2443	-0.8912
32	69	175	162.5506	1.4095	12.4494	1.5831	2.0705
33	56	154	150.6752	0.3675	3.3248	0.7477	0.8266
34	64	162	157.9831	0.4496	4.0169	1.2618	1.3004
35	36	136	132.4053	0.3960	3.5947	-0.5374	-0.2396
36	50	142	145.1942	-0.3512	-3.1942	0.3622	0.1158
37	39	120	135.1458	-1.6652	-15.1458	-0.3447	-1.1874
38	21	120	118.7030	0.1464	1.2970	-1.5013	-1.1874
39	53	158	147.9347	1.1091	10.0653	0.5549	1.0635
40	63	144	157.0696	-1.4600	-13.0696	1.1975	0.2342
41	29	130	126.0109	0.4431	3.9891	-0.9872	-0.5950
42	20	116	117.7895	-0.2025	-1.7895	-1.5655	-1.4243
43	19	124	116.8760	0.8081	7.1240	-1.6298	-0.9504
44	36	136	132.4053	0.3960	3.5947	-0.5374	-0.2396
45	50	142	145.1942	-0.3512	-3.1942	0.3622	0.1158

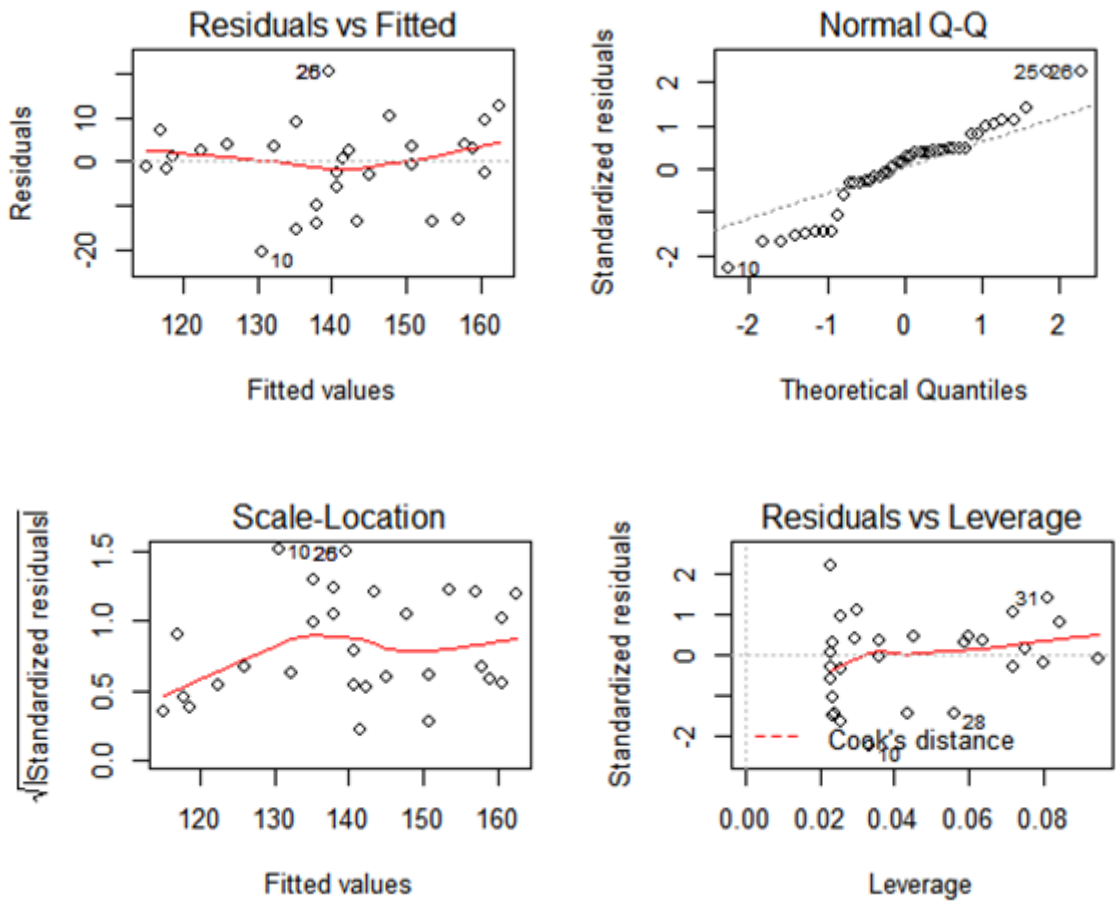


Figure2. *Scatter plot of Linear Regression models (Removed outliers). y = 0.9135x + 99.52, R² = 0.7091*

Step 1: The medical data imported from .xlsx file and attach() in R Studio and assign SBP as a dependent variable and Age as a independent variable, ie. $(X_i, Y_i)\ \ i = 1,2,..,,n$

Step 2: To get summary statistics from the given data.

Step 3: To execute nonlinear regression function with R syntax and summarize the results

Step 4: Predict the given data and get Age2 predicted values.

Step 5. To get the residuals values for given data

Step 6: To get the standardized values for given data.

Step 7. To execute scale function of linear regression for the variable Age2 and get Standardized values for given data. Scale function of each element of those values is obtained by subtracting and dividing by standard deviation

Step 8. Repeat step 7 for SBP and get the standardized values for the given data

Step 9: To Visualize the results for nonlinear regression with help of different scatter plot.

The above algorithm execute the results of nonlinear regression of Predicted value, Standardized value, Residuals, Scale function value of Age, Age^2 and SBP values. (In Table 7 and 8, Figure 3). The results table and figure show the extremes in the 14th row of the table are highlighted as outlier. In Visualization parts, models are fitted and highlighted in the scatter plots: Residuals vs, Fitted, Normal QQ plot, Scale – location and Residuals vs. Leverage. All the four plots show the 14 person as an outlier.

The summary statistics shows that the minimum age is 17 and maximum age is 69, the minimum SBP is 110 and maximum SBP as 260. The fitted linear regression model is y = 8E-04x^2+1.0015x +99.10, R^2 = 0.4899. In this model R^2 value is less than 0.5 and this model is average. Subsequently the nonlinear regression function executes the above algorithm step 1 to step 9. Cook's distance is used in

linear regression analysis to find influential outliers in a set of predictor variable. It is a way to identify points that negatively affects the regression model. The plot is a combination of each given data leverage and residual values, the higher leverage and residuals, the higher the Cook's. The F and t statistic are highly significant between two variables.

Table7. *Descriptive Statistics (n = 45)*

Regression Statistics	
R Square	0.4899
Multiple R	0.4656
Intercept	99.104
Slope (Age)	1.0015
Age ^{2}	-0.0008

R Studio Nonlinear Regression output:

Residuals:

Min 1Q Median 3Q Max

-22.228 -5.171 0.599 2.527 75.599

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 99.1047120 17.0370972 5.817 7.28e-07 ***

Age 1.0015640 0.8473475 1.182 0.244

Agw2 -0.0008047 0.0097475 -0.083 0.935

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.99 on 42 degrees of freedom

Multiple R-squared: 0.4899, Adjusted R-squared: 0.4656

F-statistic: 20.17 on 2 and 42 DF, p-value: 7.254e-07

Table9. *Nonlinear Regression Summary Output for Original Data*

S. No	Age (x_i)	SBP (y_i)	Age2	Predicted Value SBP	Standardized	Residuals	Age2 Standardized (x_i)	SBP Standardized (y_i)
1	65	162	4225	160.807	0.08349	1.19329	1.50999	0.98387
2	46	142	2116	143.474	-0.1005	-1.474	-0.0665	0.00867
3	67	170	4489	162.597	0.52733	7.40259	1.70733	1.37395
4	42	124	1764	139.751	-1.0756	-15.751	-0.3296	-0.869
5	67	158	4489	162.597	-0.3275	-4.5974	1.70733	0.78883
6	56	154	3136	152.669	0.09046	1.3311	0.69598	0.59379
7	64	162	4096	159.909	0.1453	2.09105	1.41356	0.98387
8	56	150	3136	152.669	-0.1814	-2.6689	0.69598	0.39875
9	59	140	3481	155.396	-1.0497	-15.396	0.95386	-0.0889
10	34	110	1156	132.228	-1.5156	-22.228	-0.7841	-1.5517
11	42	128	1764	139.751	-0.8024	-11.751	-0.3296	-0.674
12	48	130	2304	145.326	-1.0442	-15.326	0.07407	-0.5765
13	39	144	1521	136.942	0.48187	7.05817	-0.5112	0.10619
14	47	220	2209	144.401	5.15383	75.5993	0.00306	3.81195
15	45	138	2025	142.546	-0.3102	-4.5457	-0.1345	-0.1864
16	47	145	2209	144.401	0.04085	0.59926	0.00306	0.15495
17	65	162	4225	160.807	0.08349	1.19329	1.50999	0.98387
18	45	135	2025	142.546	-0.5149	-7.5457	-0.1345	-0.3327
19	17	114	289	115.899	-0.1393	-1.8988	-1.4321	-1.3566
20	20	116	400	118.814	-0.1994	-2.8141	-1.3492	-1.2591

21	19	124	361	117.844	0.44048	6.15605	-1.3783	-0.869
22	36	136	1296	134.118	0.12837	1.88181	-0.6794	-0.2839
23	50	142	2500	147.171	-0.3519	-5.1713	0.22058	0.00867
24	39	120	1521	136.942	-1.1566	-16.942	-0.5112	-1.0641
25	21	120	441	119.783	0.01527	0.2173	-1.3185	-1.0641
26	44	160	1936	141.616	1.25492	18.3843	-0.201	0.88635
27	44	160	1936	141.616	1.25492	18.3843	-0.201	0.88635
28	53	158	2809	149.927	0.54855	8.07267	0.45155	0.78883
29	63	144	3969	159.01	-1.0371	-15.01	1.31863	0.10619
30	29	130	841	127.473	0.17253	2.52665	-1.0195	-0.5765
31	25	125	625	123.641	0.09357	1.3591	-1.181	-0.8203
32	69	175	4761	164.382	0.77579	10.6183	1.91064	1.61775
33	56	154	3136	152.669	0.09046	1.3311	0.69598	0.59379
34	64	162	4096	159.909	0.1453	2.09105	1.41356	0.98387
35	36	136	1296	134.118	0.12837	1.88181	-0.6794	-0.2839
36	50	142	2500	147.171	-0.3519	-5.1713	0.22058	0.00867
37	39	120	1521	136.942	-1.1566	-16.942	-0.5112	-1.0641
38	21	120	441	119.783	0.01527	0.2173	-1.3185	-1.0641
39	53	158	2809	149.927	0.54855	8.07267	0.45155	0.78883
40	63	144	3969	159.01	-1.0371	-15.01	1.31863	0.10619
41	29	130	841	127.473	0.17253	2.52665	-1.0195	-0.5765
42	20	116	400	118.814	-0.1994	-2.8141	-1.3492	-1.2591
43	19	124	361	117.844	0.44048	6.15605	-1.3783	-0.869
44	36	136	1296	134.118	0.12837	1.88181	-0.6794	-0.2839
45	50	142	2500	147.171	-0.3519	-5.1713	0.22058	0.00867

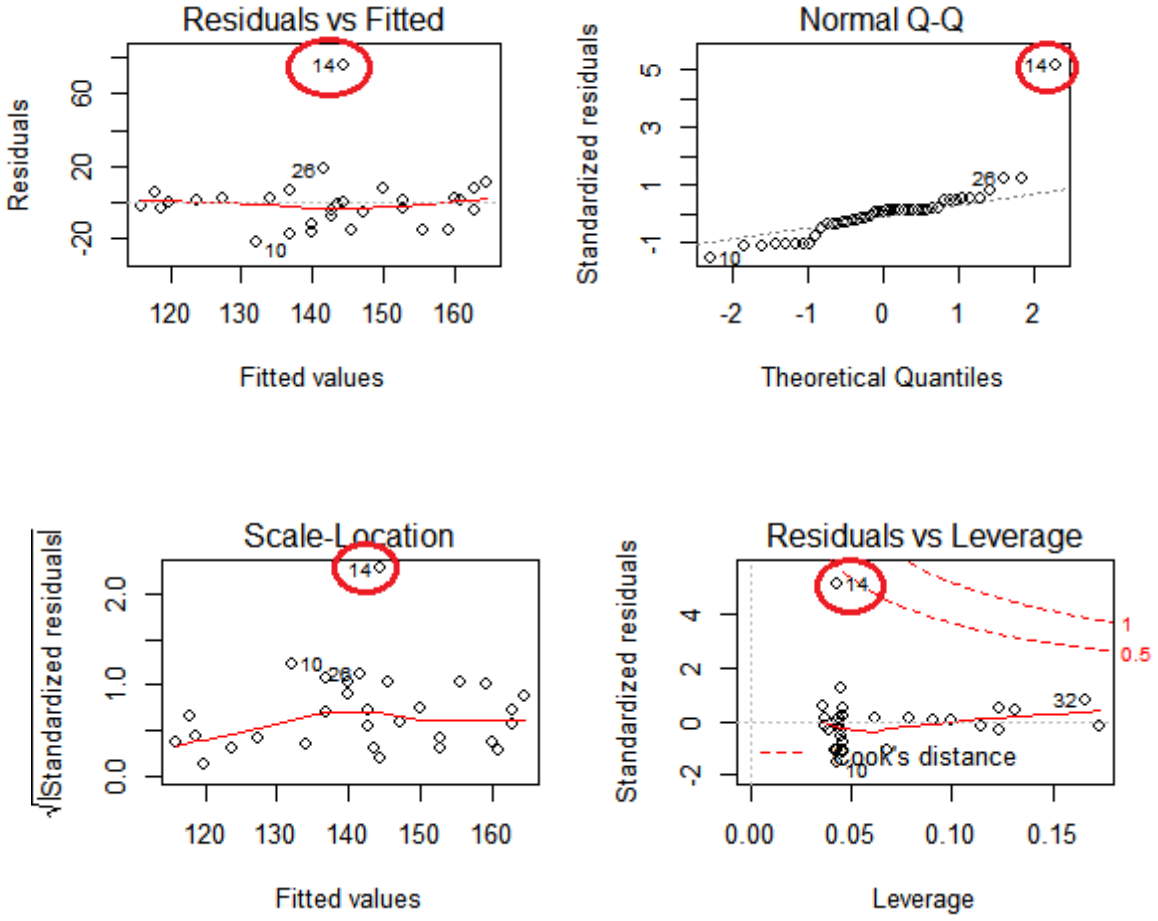


Figure3. *Scatter plot for Nonlinear Regression (with outliers) $Y=8E\text{-}04x^2+1.0015x+99.10$, $R^2 = 0.4899$*

The previous algorithm again execute the results of linear regression for outlier data and get the Predicted value, Standardized value, Residuals, Scale function value of Age and SBP values. (Table 10 to 11, Figure 4). The linear regression model improves better. In Visualization parts, models are fitted and highlighted in the plots: Residuals vs, Fitted, Normal QQ plot, Scale – location and Residuals vs. Leverage. All the four plots show better results after the removal of outliers.

The summary statistics shows that the minimum age is 17 and maximum age is 69, the minimum SBP is 110 and maximum SBP as 175. The fitted linear regression model is $y = 0.0098x^2 - 0.0363+0.109$, $R^2 = 0.7169$. In this model R^2 value is closer to .705 and this model is better. Subsequently the non linear regression executes the algorithm step 1 to step 9. Cook's distance is used in linear regression analysis to find influential outliers in a set of predictor variable. It is a way to identify points that negatively affects the regression model. The plot is a combination of each given data leverage and residual values, the higher leverage and residuals, the higher the Cook's. The F and t statistic are highly significant between two variables.

Table10. *Descriptive Statistics n = 44*

Regression Statistics	
R Square	0.7169
Multiple R	0.7031
Intercept	0.0109
Age ^{2}	0.0105

R Studio Nonlinear Regression output (Outlier Removed):

Residuals:

Min 1Q Median 3Q Max

-18.9926 -5.2332 0.6652 4.4728 22.7874

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 116.81030 2.65684 43.97 < 2e-16 ***

Age2 0.01054 0.00103 10.23 5.68e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.142 on 42 degrees of freedom

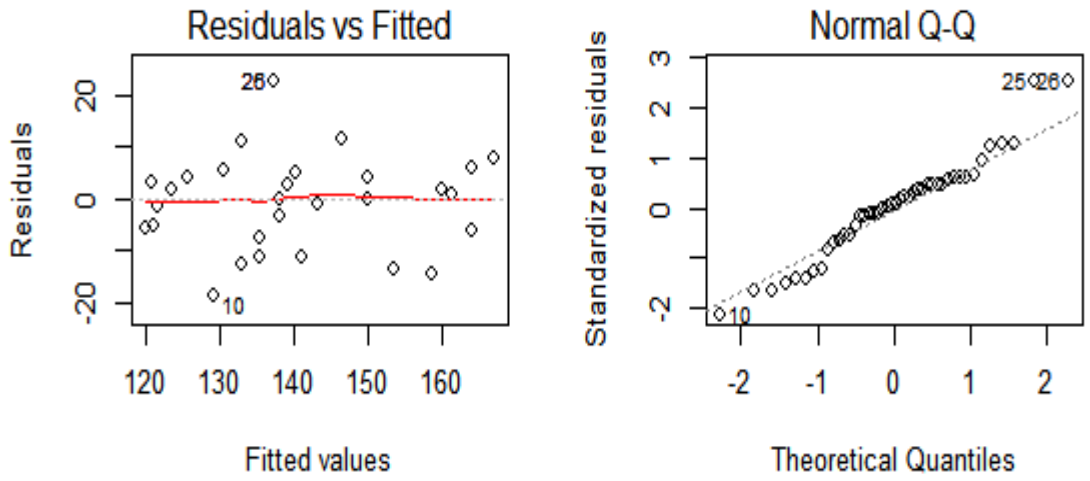
Multiple R-squared: 0.7136, Adjusted R-squared: 0.7068

F-statistic: 104.6 on 1 and 42 DF, p-value: 5.676e-13

Table11. *After removing outlier from original data (removed 14th row)*

S. No	Age (x_i)	SBP (y_i)	Age2	Predicted Value	Standardized	Residuals	Age2 Standardized (x_i)	SBP Standardized (y_i)
1	65	162	4225	161.335	0.07563	0.66516	1.32603	1.30043
2	46	142	2116	139.11	0.31985	2.89054	0.10515	0.11577
3	67	170	4489	164.117	0.67421	5.88303	1.45455	1.77429
4	42	124	1764	135.4	-1.263	-11.4	-0.1519	-0.9504
5	67	158	4489	164.117	-0.701	-6.117	1.45455	1.0635
6	56	154	3136	149.859	0.46085	4.14142	0.74772	0.82657
7	64	162	4096	159.975	0.22941	2.02461	1.26178	1.30043
8	56	150	3136	149.859	0.01574	0.14142	0.74772	0.58964
9	59	140	3481	153.494	-1.5092	-13.494	0.94049	-0.0027
10	34	110	1156	128.993	-2.1167	-18.993	-0.6659	-1.7797
11	42	128	1764	135.4	-0.8198	-7.4	-0.1519	-0.7135

12	48	130	2304	141.091	-1.2273	-11.091	0.23366	-0.595
13	39	144	1521	132.839	1.2387	11.1609	-0.3447	0.23424
15	45	138	2025	138.151	-0.0167	-0.1505	0.04089	-0.1212
16	47	145	2209	140.09	0.54334	4.91048	0.16941	0.29347
17	65	162	4225	161.335	0.07563	0.66516	1.32603	1.30043
18	45	135	2025	138.151	-0.3487	-3.1505	0.04089	-0.2989
19	17	114	289	119.856	-0.664	-5.8559	-1.7583	-1.5427
20	20	116	400	121.026	-0.5682	-5.0257	-1.5655	-1.4243
21	19	124	361	120.615	0.38314	3.38535	-1.6298	-0.9504
22	36	136	1296	130.468	0.61542	5.53199	-0.5374	-0.2396
23	50	142	2500	143.156	-0.128	-1.1562	0.36218	0.11577
24	39	120	1521	132.839	-1.425	-12.839	-0.3447	-1.1874
25	21	120	441	121.458	-0.1647	-1.4577	-1.5013	-1.1874
26	44	160	1936	137.213	2.52259	22.7875	-0.0234	1.18196
27	44	160	1936	137.213	2.52259	22.7875	-0.0234	1.18196
28	53	158	2809	146.413	1.28519	11.5875	0.55495	1.0635
29	63	144	3969	158.637	-1.6533	-14.637	1.19752	0.23424
30	29	130	841	125.673	0.48466	4.32694	-0.9872	-0.595
31	25	125	625	123.397	0.18034	1.60323	-1.2443	-0.8912
32	69	175	4761	166.983	0.92726	8.0166	1.58306	2.07045
33	56	154	3136	149.859	0.46085	4.14142	0.74772	0.82657
34	64	162	4096	159.975	0.22941	2.02461	1.26178	1.30043
35	36	136	1296	130.468	0.61542	5.53199	-0.5374	-0.2396
36	50	142	2500	143.156	-0.128	-1.1562	0.36218	0.11577
37	39	120	1521	132.839	-1.425	-12.839	-0.3447	-1.1874
38	21	120	441	121.458	-0.1647	-1.4577	-1.5013	-1.1874
39	53	158	2809	146.413	1.28519	11.5875	0.55495	1.0635
40	63	144	3969	158.637	-1.6533	-14.637	1.19752	0.23424
41	29	130	841	125.673	0.48466	4.32694	-0.9872	-0.595
42	20	116	400	121.026	-0.5682	-5.0257	-1.5655	-1.4243
43	19	124	361	120.615	0.38314	3.38535	-1.6298	-0.9504
44	36	136	1296	130.468	0.61542	5.53199	-0.5374	-0.2396
45	50	142	2500	143.156	-0.128	-1.1562	0.36218	0.11577



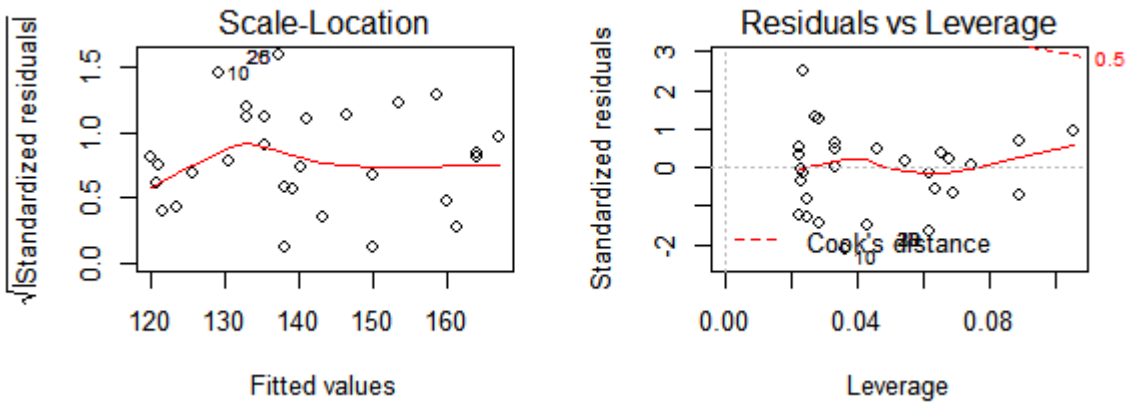


Figure4. *Scatter plot for Nonlinear Regression (Outlier Removed)* $y = -2E\text{-}06x^2 + 0.0183x + 112.2$, $R^2 = 0.4904$

5. CONCLUSION

Linear and non linear regression analysis assumes scatter of data, fitting of straight line or normal distribution. An outlier is an extreme observation when the residual is larger in absolute value when compared with the other observed data set. The detection of outlier can be defined as the process of detecting and subsequently excluding outliers from the given set of data. Outlier can dominate the sum of the square calculation and might lead to misleading results.

The linear and nonlinear regression model fitted for original and outlier removed data. The results of original data linear and non linear regression are not a good model. In both model R^2 value is less than 0.5. After removal of the outlier, it achieved better fit of linear and nonlinear regression model. The R^2 values are more than 0.7. The F and t statistic are significant in two models. The scatter plot clearly visualizes the outlier and without outlier data for different plots. In this paper, the detection of outliers in simple linear regression model has been discussed. A new approach for detecting outliers without the use of predicted values have been proposed, which is quite useful in detecting outliers that detects the outliers as same as the residual and standardized residual method. Hence, it is proposed, that in simple linear regression model, the difference method can be used for detecting outliers. In general the linear and nonlinear regression model is used to removal of outlier for any given data set.

REFERENCES

[1] N. R. Draper, Norman Richard Draper, Harry Smith (1981), Applied Regression Analysis, Third Edition, A John Wiley & Sons, Inc., Publication, New Jersey, USA.

[2] Framstad, Erik, Steinar Engen, and Nils Chr. "Regression analysis, residual analysis and missing variables in regression models." Oikos (1985): 319-323.

[3] Manimannan G. and R. Lakshmi Priya (2019), "Evaluation and Classification of Master Health Checkup Database using Data mining Techniques", International Journal of Data Mining and Emerging Technologies, Vol.9, Isssue:2 pp. 25-32.

[4] Douglas Montgomery, *et. al.* (2012) Introduction to Linear Regression Analysis, Fifth Edition, A John Wiley & Sons, Inc., Publication, New Jersey, USA.

[5] Bipin Gogoi and Mintu Kr. Das. Usage of graphical displays to detect outlying observations in linear regression. Indian Journal of Applied Research 5.5 (2015): 19-24.

[6] Hampel, F.R. *et.al* (1986), Robust Statistics: The Approach Based on Influence Functions. New York: John Wiley & Sons, Inc.

Citation: *Manimannan G,et.al., Detecting Outliers using R Package in Fitting Data with Linear and Nonlinear Regression Models, International Journal of Scientific and Innovative Mathematical Research (IJSIMR), vol. 8, no. 4, pp. 1-13, 2020. Available : DOI: https://doi.org/10.20431/2347-3142.0804001*